ELSEVIER

The Journal of Socio-Economics 33 (2004) 665–675

The Journal of Socio-Economics

www.elsevier.com/locate/econbase

Significance redux

Stephen T. Ziliak[a,*], Deirdre N. McCloskey[b,1]

[a] *School of Policy Studies, College of Arts and Sciences, Roosevelt University, Chicago, 430 S. Michigan Avenue, Chicago, IL 60605, USA*

[b] *College of Arts and Science (m/c 198), University of Illinois at Chicago, 601 South Morgan, Chicago, IL 60607-7104, USA*

Abstract

Leading theorists and econometricians agree with our two main points: first, that economic significance usually has nothing to do with *statistical* significance and, second, that a supermajority of economists do not explore *economic* significance in their research. The agreement from Arrow to Zellner on the two main points should by itself change research practice. This paper replies to our critics, showing again that economic significance is what science and citizens want and need.

JEL CODES: C12; C10; B23; A20

Keywords: Hypothesis testing; Statistical significance; Economic significance

Science depends on entrepreneurship, and we thank Morris Altman for his. The symposium he has sparked can be important for the future of economics, in showing that the best economists and econometricians seek, after all, *economic* significance. Kenneth Arrow as early as 1959 dismissed mechanical tests of statistical significance in his search for *economic* significance. It took courage: Arrow's teacher, the amazing Harold Hotelling, had been one of Fisher's sharpest disciples. Now Arrow is joined by Clive Granger, Graham

* Corresponding author.

E-mail addresses: sziliak@roosevelt.edu (S.T. Ziliak), deirdre2@uic.edu (D.N. McCloskey).

[1] For comments early and late we thank Ted Anderson, Danny Boston, Robert Chirinko, Ronald Coase, Stephen Cullenberg, Marc Gaudry, John Harvey, David Hendry, Stefan Hersh, Robert Higgs, Jack Hirshleifer, Daniel Klein, June Lapidus, David Ruccio, Jeff Simonoff, Gary Solon, John Smutniak, Diana Strassman, Andrew Trigg, and Jim Ziliak.

1053-5357/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.socec.2004.09.038

Elliott, Joel Horowitz, Ed Leamer, Tony O'Brien, Erik Thorbecke, Jeffrey Wooldridge, and Arnold Zellner (Arnold is in our minds a Zeus in the matter of economic significance, with Ed Leamer as his Mercury). Their examples inspire hope. And our hope is strengthened learning as we do here from colleagues in cognate fields that mechanical tests have been criticized by their best, for decades. It's time to stop the nonsense and get serious about significance in economics.

Why has it taken until now for economists to catch on? In his own paper Morris Altman makes a good case for path dependence. People have believed that mechanical testing for statistical significance is all right because, after all, it's been around for so long, something one might say, too, of the labor theory of value, or protectionism, or belief in séances with the dear departed. As Altman observes, even in psychology, where since the Significance Test Controversy of the early 1970s there has been widespread understanding of the issue by sophisticates, little has changed. Fidler et al. conclude here, too: "psychology has produced a mass of literature criticizing null-hypothesis statistical testing over the past five decades. . . but there has been little improvement. . .. Even editorial policy and (admittedly half-hearted) interventions by the American Psychological Association have failed to inspire any substantial change." Capraro and Capraro (2004), cited in Altman, found that in psychology the number of pages in texts and guidebooks recommending the mechanical use of statistical significance was orders of magnitude larger than the number of pages warning that after all effect size is always the chief scientific issue. Our papers show the same to be true for a supermajority of econometrics texts, from the advanced *Handbook of Econometrics* through Arthur Goldberger's latest down to the simplest of introductory textbooks. Students get misled from the beginning. Few see a problem. And even fewer break away.

But the present forum may be the beginning of the end for a silly and unscientific custom in economics. We associate ourselves with the remark by the psychologist W.W. Rozeboom in 1997, quoted by Bruce Thompson here (Rozeboom has been making the point since 1960): "Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students. . .. It is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism" (Rozeboom, in B. Thompson, p. 335). Precisely.

1. How to deal with random error

To unblock the journal referees and editors and break out of what Altman calls "a steady-state low-level equilibrium" we propose asking major economists and econometricians to state publicly their support for the following propositions: (1) Economists should prefer confidence intervals to other methods of reporting sampling variance. (2) Sampling variance is sometimes interesting, but is not the same thing as scientific importance. (3) Economic significance is the chief scientific issue in economics; an arbitrary level of sampling significance is no substitute for it. (4) Fit is not a good all-purpose measure of scientific validity, and should be deemphasized in favor of inquiry into other measures of importance. Every editor of every major journal will be asked. We think that on reflection most economists and econometricians will agree with these propositions.

Scores of the best statistical investigators in psychology, sociology, and statistics itself have been making such points for a long time, longer even than McCloskey has, who came by her insights honestly, stealing them fair and square 20 years ago from pioneers like Denton Morrison and Ramon Henkel in sociology and Paul Meehl and David Bakan in psychology and Kenneth Arrow in economics (Arrow, 1959; McCloskey and Ziliak, 1996). Ziliak first learned about the difference between economic and statistical significance in the late 1980s, when he purchased for his job at the State of Indiana Department of Employment and Training Services an elementary book by the two Wonnacott brothers, one an economist, the other a statistician (Wonnacott and Wonnacott, 1982, p. 160). But he met puzzling resistance when he argued to the chief in his division of labor market statistics how it was a shame that rates of unemployment among black urban teenagers in Indiana were not being published, and were therefore not being discussed openly and scientifically, merely because their small sample sizes did not attain conventional levels of statistical significance.

Good fit modulo the present "sample" is nice, even "neat." But there is no reason to make fit the criterion of model selection. As Arnold Zellner points out in his comments, sometimes of course the fit measured by R^2 is perfect because the investigator has regressed US national income precisely on itself. Y fits Y, L fits L, K fits K. "Fit" in a wider scientific sense, which cannot be brought solely and conveniently under the lamppost of sampling theory, is more to the point. How well for example does the model (or parameter estimate) fit phenomena elsewhere? Are there entirely different sorts of evidence—experimental, historical, anecdotal, narrative, and formal—that tend to confirm it? Does it accord with careful introspections about ourselves? What could be lost if policymakers or citizens act as if the hypothesis were true? So we remain skeptical that some simple and equally mechanical refinement of statistical significance will work. Some of the advanced proposals miss the main point, that fit is not the same thing as importance.

2. Precision is nice but oomph is the bomb

The kind of decision-making we advocate can be illustrated thus. Suppose you want to help your mother lose weight, and are considering two diet pills of identical price and side effects. The one, named "Oomph," will on average take off 10 pounds, but it is rather uncertain in its effects, at plus or minus 5 pounds. Not bad. Alternatively, the pill "Precision" will take off only 3 pounds on average, but it's less of a roll of the dice: Precision brings a probable error of plus or minus 1 pound. How nice.

The signal-to-noise ratio of diet pill Oomph is 2:1, that for Precision 3:1. Which pill for Mother? "Well," say some of our scientific colleagues, "the one with the highest signal-to-noise ratio is Precision. So, of course: hurrah for Precision." Wrong, of course; wrong, that is, for Mother's weight-management program and wrong for the distressingly numerous victims of scientists in the misled fields from medicine to management. Such scientists decide whether something is important or not, whether it has an effect, as they say, by looking not at its oomph but at how precisely it is estimated. But the pill Oomph promises to shed 5 or 15 pounds. The much less effective Precision will shed less than 4 pounds. Common sense recommends Oomph. The burden of this symposium is: let's get back to common sense—to oomph—in science.

The crucial thing to grasp in the comments gathered here is this: *every one of the commentators agrees with our two main points*:

1. that economic significance usually has nothing to do with *statistical* significance, and
2. that a supermajority of economists do not explore *economic* significance in their research.

The agreement from Arrow to Zellner on the main points should by itself change research practice. Moreover, the tiny objections the critics raise against us, though significant as sociology of science, in no way undermine the consensus. Economic significance, substantive significance, is the body, not statistical significance unadorned. We all here agree.

3. Some reasons statistical significance does not select models

Graham Elliott and Clive Granger agree with our point, but want for some reason to characterize it as "literary" and not "deep." Perhaps it arises from their mistaken belief that if sample means and so forth are somewhere provided in a paper, then "the economic significance can be determined." Set aside that, as they admit, in many cases the papers do not provide the data to get beyond a statement that a certain coefficient is or is not "significant." Our main point is not this stylistic one. It is that "significance" itself is something that needs to be argued out in the context of the scientific or policy issue and cannot be determined on statistical grounds alone. Our point is not to repeat a matter of style, literary matters, or superficialities of presentation. The economic significance cannot "be determined" by simply better reporting on conventional statistical tests. The mistake of Elliott and Granger shows in their claim that what would be at issue in cases of bad reporting is the "statistical comprehension skills" of the reader. No. It is the *economic* comprehension skills that matter for economic science: that is our main point. We cannot hand science over to a table of Student's *t*.

We have learned recently, by the way, that "Student" himself—William Sealy Gosset—did not rely on Student's *t* in his own work. To the world's gain Gosset's job and passion was to instead learn scientifically how to brew the best Guinness he could brew at the best price the market could bear (see for example E.S. Pearson, '*Student*': *A Statistical Biography of William Sealy Gosset* [Oxford: Clarendon Press, 1990, pp. 20, 30–31]). Student used his *t*-tables a teensy bit; but Student gave his scientific time and consideration to proportions of yeast and mash, mixing ingredients over time for a maximum oomph in Guinness, as you'd want and expect. R.A. Fisher begged Student for his tables of *t* to publish in Fisher's now hugely damaging *Statistical Methods for Research Workers*. Yet Fisher—himself a decent farmer—did not as we have shown believe he needed to emulate Student's care for *magnitudes* of ingredient effect, and focused instead on *t*.

Often we focus on how to interpret the parameters of a specific model. Elliott and Granger agree with us but then focus their critical energies on a defense of mechanically computed statistical significance to separate theory A and theory B (we believe they mean "model" A and model B, though their comments equivocate.) We are not persuaded.

Their instance in physics, of the large, Einsteinean bending of light around the sun as against the Newtonian prediction of less bending, is ill chosen. The physicists making the experiment did not in fact use statistical tests. The leader of the historic 1919 expeditions

to photograph the eclipsed sun off the coasts of West Africa and Northern Brazil (to see the bending light to which Granger and Elliott refer) was Sir Arthur Eddington, the Cambridge astronomer and popularizer of relativity. Eddington, it turns out, had been a teacher of the statistician Harold Jeffreys, and Jeffreys was intensely interested in the results of the expedition. Arnold Zellner has tried with little avail for decades to get economists to read Jeffreys, precisely because Jeffreys believed, against his teacher Sir Arthur, that statements of "existence" are for purposes of hypothesis and model testing useless (Wrinch and Jeffreys, 1921; Howie, 2002, pp. 92–3; Ziliak and McCloskey, this volume). Size is what mattered in the Einstein–Newton debate; size always matters. The photographic evidence was not at first persuasive; indeed, it is well known in the history of science that it was some years before an error caused by the instrumentation was corrected: Einstein's theory was at first rejected by the evidence. And so Eddington reasoned in favor of Einstein on geometric, a priori grounds. Jeffreys (whom we also highly recommend) and his collaborator Dorothy Wrinch responded with an empirically based criticism of Eddington's defense, and published their piece in a now famous issue of *Nature* in which Einstein considered all the evidence (Wrinch and Jeffreys, 1921). Size, instrumentation, design of sample, varied observations, coherence with other stories and other kinds of evidence are what persuaded. No tests of statistical significance, Jeffreys and Einstein agreed, could alone shed light.

Indeed, as we report, in the leading journals of physics such as the *Physical Review* one hardly ever encounters the t, p, R^2, and the like that litter journals of economics, psychology, and medicine. Physicists certainly do test one physical model A against a rival B. But they never hand over the criterion of decision to an unargued level of significance. Ask any physicist. One of us last month for example asked a distinguished physicist who was helping out with the selection of Phi Beta Kappa Awards. Roughly he said in reply, "Of course not. We use statistical models, such as Brownian motion. But never do we 'test' at arbitrary levels of significance the way biologists sometimes do."

No wonder. Suppose you were comparing two pieces of silverware, one a spoon, A, and the other a fork, B. Suppose you wanted to know how similar A was to B. The procedures we and the numerous other critics in other fields are complaining about are mechanical "tests" *on the half-inch of pattern on the "handles" of each piece*. The comparison of models is reduced to the comparison of fit in the so-called "sample" on offer. These may turn out to be very similar—imagine the spoon and the fork coming from the same silverware pattern, and so having much the same figuration of the end of the handles. But a fork in its forked end is different from a spoon in its spooned end for use, for science, for policy. You can't stab meat with a spoon. And no amount of mistaken reports on the philosophy of science will induce a thin soup to pool upon your fork. Precision does not pick the model. Oomph, and the scientist's stories about oomph, does.

Elliott and Granger take the view that conventional statistical methods simply *are* the techniques of "empirical methods in all of science." This is factually mistaken, though rather typical of the way statistical methods is taught these days in economics—all handles, and egg on the face. When we, and Elliott and Granger, criticize for example the mechanical use of 5% significance levels we are criticizing a practice that is widespread only in a tiny part of science. Though it is a part that Clive Granger could singlehandedly reform if he would!

We are surprised that our old friend Joel Horowitz, who we know agrees with much of what we say, asserts "there are circumstances in which the existence of a phenomenon, not its magnitude, is decisive." Horowitz, unlike us, was trained as a physicist. But here he is talking like a mathematician. We prefer the talk of physicists, such as Richard Feynman, whose great "elementary" textbook at Cal Tech is filled with statements like "are zero, *or can be neglected in comparison with the variations in the other directions*" (II, p. 7–2) or "the fact that there is an amplitude ... *has little effect* when the two positions have very different energies" (III, p. 9–8). Or in his lectures on computation that "Predictive coding enables us to compress messages to *a quite remarkable degree*" (1984–86 [1996], p. 129). Horowitz will be able to tell us what on earth Feynman was talking about so far as the physics is concerned. But what is obvious in Feynman's talk even to an outsider is that it is about *magnitudes*, never about existence in the mathematician's sense. Remember from your math course: a mathematician trying to prove that a number is greater than zero doesn't care a fig whether the number is 10^{100} or 10^{-100}. The physicist does, every time. All right: nearly every time. Thus the famous case of Feynman's test with a glass of water during the Challenger inquiry: was a temperature around freezing *low enough* to change the behavior of the stuff used for the O rings *low enough to matter*?

We see the point of Horowitz's example of Cronin and Fitch. But presumably if the effect had turned out to be two orders of magnitude greater than it was in fact, then the surrounding physics would have been greatly altered. So magnitude mattered even in that case of a very faint effect. And as he himself says, economics is not precise enough for tiny effects to be relevant anyway, a point made 50 years ago by Oskar Morgenstern. The problem with Horowitz's "existence" talk—which, we repeat, we do not think even he believes is very important, since on the whole he agrees with us and teaches our point to his students—is that it suggests there must be a "test" for it, *free of any worries about how big is big*. But there isn't. When Horowitz says that "the difference between [0.2 and 0.4]... is interesting and important only if we can be reasonably sure that it is not an artifact of random sampling error"he is applying an arbitrary criterion of statistical significance, which after all is the main thing both he and we don't like. The point is that *even if* (say) *a* 95 *percent confidence interval contains both* 0.2 *and* 0.4 that doesn't mean there "exists no difference," or that we are justified in thinking there is "no difference" in the predictions of the two theories (say). *It depends on the loss function*. To put it another way, it depends on the significance level one chooses relative to alternative hypotheses, and even that (as Neyman and Pearson stressed) is a scientific and social decision, not to be left to convention or ritual disguised as a mere formality. After the recent one hundred years of economic growth the difference between a well-fitted 1.1% annual average rate of growth in real GDP and a well-fitted 2.2% annual average rate of growth in real GDP is the difference between Argentina and France, where income per person now differs by some $16,000. If one just had to make from the Solon data some crucial decision, and had got a coefficient of 0.4, though alas from very noisy data, one might have to go ahead and suppose that 0.4 was The Truth.

We do not wish with Quetelet to take random error out of economic or physical or other accounts of the world. Noise exists, and sometimes one wants to know how much there is, and distinguish it from some effect of actual interest. Fine. But we are sure Horowitz would agree that *this does not justify using statistical significance to decide on what variables are*"*important*," which as we have shown is the usual economic practice. In fact the coeffi-

cient of 0.4 in question, from Gary Solon's 1992 study of intergenerational income mobility, passed conventional tests and seems moreover to be the product of a Pareto improved model for extracting income parameters.

We agree with the more radical point of another old friend of ours, Ed Leamer, that economics needs tests of persuasiveness or usefulness, both of which could be called in official philosophical language "Pragmatism." Leamer is correct that tests of significance persist precisely because they do not in fact settle much that persuades scientists intent on usefulness. Consider the enormous number of tests of significance done each year on both sides of every issue in economics. Would it surprise anyone to assert that they were, let us say, on the order of 10 million? If the tests were in fact as conclusive as their own rhetoric requires, most issues in economics would long have been settled. That's one way of putting Leamer's point. We see some similarity here between Leamer's point and the very interesting argument with which Horowitz ends his paper. In any event we are confident that both Leamer and Horowitz would agree with us that when one wants compare a spoon and a fork perhaps it would be wise to develop other ways of comparing them beyond doing statistical tests on the design of the handles over and over and over again, 10 million times, to no one's enlightenment.

We agree with Peter Lunt that R.A. Fisher's intent in the 1920s and for a long time after was "to develop what he hoped would be conventionally agreed, automatic procedures for statistical inference," because judgment is "fallible." We agree by the way with the spirit of Gigerenzer's paper but we do not agree with his reading of Fisher: Fisher, as we show, "invents" the "formal" criterion of statistical significance, for sure by 1925, and doubtless somewhat earlier in personal communications. And Fisher's less well known second thoughts on the matter, that is, his coming to believe correctly that a "rule of 2" and the like is foolish were we think wariness caused by a stronger mathematician, Jerzy Neyman, who showed Fisher the clumsiness of Fisher's qualitative, yes/no reasoning and the incompleteness of his approach to error, which was Type I only. Fisher therefore took late in life the Jeffreys inspired "science route"—saying finally correctly that a mechanical rule is silly and that what scientists really care about is magnitudes of effects and relations, not mathematical "existence."

But Lunt understands us to be Fisherians, wishing only to improve the practice that came from Fisher's temporary victory in practice over Neyman and Pearson. That is mistaken: we are Neymanites, and Jeffreysites, and most assuredly are not modernist positivists (see for example McCloskey, 1983,1985 [1998], 1990, 1994; Ziliak, 2001, 2003, 2005). We would be very willing to engage in an epistemological critique of economics, and in fact we have opened that particular Pandora's Box on many occasions, and at length. But on this occasion we are engaging, as Lunt says, in an "internal critique." It seems appropriate. If economists can't get even their mechanical methods right, perhaps they need to consider a broader range of ways of arguing—for example (to again stay within conventional economic method) putting more emphasis on the simulation that has been made so easy by the fall in computation costs. On the other hand, we agree with Lunt that the analogy of regression analysis with experimental method on which classical econometrics is built may be reaching a crisis.

We agree with Tony O'Brien that the next step is to see how badly economists are doing in their subfields, such as economic history, O'Brien's target, in making this childish mistake

in statistical procedure. O'Brien's project is harder to do, of course, because one has to get down deeper into the discourse, to see how the evidence and argument are constructed overall. That is, one needs to see how the rhetoric works. O'Brien believes that childish mistakes do not always have bad consequences. We agree with him that the character of research in economic history keeps the results from depending too much on the mistakes. Exactly as he says, if one really knows a subject one wants to know about oomph. (And we are proud to say that economic historians generally know their subjects much better than do economists satisfied with manipulating the same old one-in-a-thousand Michigan samples over and over again, or the same old quarterly time series over and over again.) We believe indeed, as O'Brien appears to find inconceivable, that economic historians in fact do "their economics better than the authors in the *AER*." The economic historians published in the *AER* in fact score much higher on our questionnaire than, say, the average macro or finance economist.

4. Defineability of economic significance

Erik Thorbecke does the best job of summarizing our paper in his own words, which makes us think that he grasps it the best. Richard Feynman used to say that if you cannot express your physical argument clearly enough to give a lecture on it to undergraduates you don't really understand it. Thorbecke's diagram in particular is brilliantly illuminating. Like O'Brien, however, Thorbecke is not sure how (or how we want) to proceed. His desire for concreteness, for more examples of Standards of Economic Significance, is understandable.

We recommend starting with magnitudes already discovered and offered as "standards" of economic significance. Often the magnitudes will serve as a kind of reservation price, indicating whether the thrust of a story or a policy ought to shift, and how. A leading example is the yeoman work of Robert William Fogel estimating the social savings attributable to the railway. Other examples abound. Gary Solon has pushed the quantitative side of the left's argument about social immobility to a new and challenging level. Any scientist or historian intending to answer the question of social or income class mobility in the United States must at some point confront Solon's estimates. Solon's story says now that sons are strongly fated by their fathers. Sometimes an examination of simple statistics over a patch of neglected history can help to establish a context for economic significance. Take, for example, the 10^n statistical studies of welfare reform. The 1996 Personal Responsibility and Work Opportunity Reconciliation Act was passed in part on the belief that the welfare state had enabled poor people to stay on welfare for life and therefore a great many did. But the historical record shows in fact that the average length of time a family stays on relief, private or public, religious or secular, has not much changed since the 1820s—families stay on relief on average not for a lifetime but for 8–13 months (Ziliak, 2002, 2004). Still, the journals continue to fill with papers reporting uselessly "significant" results on welfare checks and the duration of welfare dependence, ignoring history and its counterfactual coefficients.

Against much lower estimates put forth by the Council of Economic Advisors (CEA), James P. Ziliak, et al. have argued that up to 75% of the 1990s decline in welfare participation was caused by improvements in economic growth (Ziliak et al., 2000). The White House

claimed that the new, draconian welfare laws were the main cause of the decline, citing the lower figures the CEA attributed to economic growth. Complete agreement on the quantitative contribution of economic growth to the decline in welfare rolls may not ever emerge. But the economists in the debate understand that the size of economic and social magnitudes and their overall effects on outcomes and decision-making are what is at stake. Not anyway, statistical significance, which is of course amply supplied by each side. Every macroeconomist is familiar with the standards of economic significance exemplified by "Taylor's rule," the sacrifice ratio, and "Okun's law." The main point of economic science, as Thorbecke clearly sees, is to discover the magnitudes of relations between economic variables and then argue them out.

Jeffrey Wooldridge, like Thorbecke and all the critics here, gets the point. Wooldridge is indeed a champion of economic significance—a fact we hope is common knowledge among the young econometricians who strive to emulate him. Says Wooldridge, "I attend too many empirical workshops where the sizes of the coefficients are not discussed"—which is simply more evidence, he believes, "that econometric practice may indeed be in trouble." Yes, indeed. We do not agree however with his claim that we "oversell" the extent and error of sign and asterisk econometrics. Wooldridge cites the Bernheim and Wantz paper of 1985—a paper we say exemplifies the very problems—and lets them off the hook on grounds that "while the coefficients have the signs they expect from theory, they are not willing to claim additional support for their theory because the effects are statistically insignificant." The problem we see—and we believe Wooldridge will on reflection agree with us—is that claiming additional support for the negative signs because the effects are statistically significant is by this logic of off-the-hook equally valid. Yet neither claim is valid. A tightly fit and negatively signed coefficient on bond yield may be for economic purposes zero and insignificant. As we've argued at length, and as Wooldridge suggests in his excellent textbook, sign without size, and sign without size without confidence intervals, and sign without size without confidence intervals without loss functions, is mainly beside the point. The sign laid bare, below its ancient flickering star, has fixed the gaze of many an economist, but never have such symbols revealed arguments or magnitudes of economic relevance.

Wooldridge agrees strongly with our claim that statistical significance is not a necessary or sufficient condition for economic significance. He is certainly correct to caution however that some researchers push magnitudes of economic significance in ways they should not. His is an assertion about the ethics of communication in science and public affairs that we find both poignant and understudied—especially in an era normalizing the so-called "reality programs" and shock jocks. But "pushing" an economically large though noisily estimated effect may not be a misuse—or "stretch," as he says—of professional stature; it may be precisely the ethical thing to do. As we showed in our 1996 paper, the noisily estimated benefit-cost ratio of 4:1 in the State of Illinois unemployment insurance program is one such instance—though lost, it seems, by followers of R.A. Fisher. The loss of jobs and wages attendant to the action, *no change in employment policy*—which is what the mechanical rule of statistical significance suggested—we find appalling, no "stretch" at all. A similar point could be made about the failure of the Labor Department to release certain black urban unemployment rates.

Reasons for acting upon a large and economically significant effect that is not statistically significant (or not yet statistically significant, one should say) are especially clear in medicine, where the outcome is sometimes life or death. In the 1970s, when the null hypothesis testing ritual was first beginning to take serious hold in journals such as the *New England Journal of Medicine*, one finds patients in control groups falling seriously ill or dead. "Of course," one replies, "people die every day." But the preventable sickness or death was often caused by placebo, by lack of treatment, the doctors nearly admitted. The placebo "control group"—the killing—would not be stopped however on grounds that for example at $N =$ 'less than 30' they had not yet found statistical significance. It is not possible to focus "too much" on economic, or human, significance. It is possible to look at the wrong magnitudes, or at the wrong samples. It is possible to design immorally an experimental control group, and to push those magnitudes. And it is possible for a time to turn a small truth into a larger one, as Wooldridge argues, and we agree, is now happening in deployments of instrumental variables and two stage least squares. But as Martin Luther King, Jr. used to say, after Carlyle, "no lie can live forever."

We were pleased to find that Arnold Zellner agrees with what we say. His own critique of practice cuts deeper than ours. We honor his and Leamer's Bayesian approach, and note his friendly and non-ideological invitation to classicists to find the unity of the two approaches. If economists did as he has been recommending for decades, testing and estimation would change immensely. Economic research would be about the measurement and meaning of the size of economic effects and economists of all persuasions, experimental and observational alike would, like Zellner and his mentor Harold Jeffreys, become much more "humble." No longer would it be possible for an editor of the *AER*, or any other journal, to force an author to use tests of statistical significance instead of likelihood ratios, as one such editor did to Jack Hirshleifer and co-authors Vernon Smith and Yvonne Durham (Hirshleifer, 2004)—unsurprisingly, Hirshleifer, Smith, and Durham found that the tests they were forced to make and report rejected at the 1% level any null hypothesis of economic interest. *Dear Diary*: *But it is Science*, *isn't it*?

William James in 1907 noted the "classical stages of a theory's career. First, you know, a new theory is attacked as absurd; then it is admitted to be true, but obvious and insignificant; finally it is seen to be so important that its adversaries claim that they themselves discovered it." We certainly hope so.

References

Arrow, K. 1959. "Decision theory and the choice of a level of significance for the *t*-test." In: Ingram Olkin, et al. (Eds.). Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford, California: Stanford University Press.

Hirshleifer, J. 2004. Personal communication with McCloskey and Ziliak. University of California, Los Angeles.

Howie, D. 2002. Interpreting Probability: Controversies and Developments in the Early Twentieth Century. Cambridge University Press, Cambridge.

McCloskey, D.N., 1983. The Rhetoric of Economics. Journal of Economic Literature 31, 482–517.

McCloskey, D.N. 1985. (1998). The Rhetoric of Economics, second ed. University of Wisconsin Press, Madison.

McCloskey, D.N., 1990. If You 're So Smart: The Narrative of Economic Expertise. University of Chicago Press, Chicago.

McCloskey, D.N., 1994. Knowledge and Persuasion in Economics. Cambridge University Press, Cambridge.
McCloskey, D.N., S.T. Ziliak, 1996. The Standard Error of Regressions. Journal of Economic Literature XXXIV (March), pp. 97–114.
Wonnacott, R.J. and T.H. Wonnacott, 1982. Statistics: Discovering Its Power. John Wiley & Sons, New York.
Wrinch, D., Jeffreys, H., 1921. The Relationship Between Geometry and Einstein's Theory of Gravitation. Nature 106, 806–809.
Ziliak, J.P., Figlio, D.N., Davis, E.E., Connolly, L.S., 2000. Accounting for the decline in AFDC caseloads: welfare reform or the economy? Journal of Human Resources 35 (3), 570–586.
Ziliak, S.T. (Ed.), 2001. Measurement and Meaning in Economics: The Essential Deirdre McCloskey. Edward Elgar Ltd., Cheltanham, UK.
Ziliak, S.T. (Ed.), 2002. "Some tendencies of social welfare and the problem of interpretation." Cato Journal 21 (3, Winter), 499–513.
Ziliak, S.T. (Ed.), 2003. "Freedom to exchange and the rhetoric of economic correctness." In: Samuels, W.J., Biddle, J.E. (Eds.). Research in the History of Economic Thought and Methodology 21-A. Elsevier Press, Amsterdam, pp. 331–341.
Ziliak, S.T., 2004. Self-reliance before the welfare state: evidence from the Charity Organization Movement. Journal of Economic History 64 (2 June), 433–461.
Ziliak, S.T. (Ed.), 2005. (forthcoming). "Interpretative econometrics and the resurrection of economic significance." In: Garnett Jr., R., Harvey J.T. (Eds.). Heterodox Economics. University of Michigan Press, Ann Arbor.